Version 1

NASA Earth Science Data Preservation Content Specification

Status of this Document

This document provides information to the NASA Earth Science Division (ESD). This document specifies contents of data and associated documentation that must be preserved from NASA Earth Science missions to enable future understanding and use of the data products they generate. Distribution of this document is unlimited.

Change Explanation

N/A

Abstract

This document defines the contents of data, metadata and associated documentation to be preserved beyond the life of missions funded by NASA's Earth Science Division. The purpose of the document is to identify all the content items that need to be preserved to ensure their availability to support future investigations in long-term scientific research. The focus of this document is on the contents (i.e., "what") and not on the implementation or representation (i.e., "how") of the content items. The content items are divided into eight categories: Preflight/Pre-Operations Calibration, Products (Data), Product Documentation, Mission Calibration, Product Software, Algorithm Input, Validation and Software Tools. Items are described under each of these categories along with rationale for requiring their preservation.

Table of Contents

INTRODUCTION

One of NASA's strategic objectives is to "Study Earth from space to advance scientific understanding and meet societal needs". NASA's Earth Science Data System (ESDS) program resides within NASA's Earth Science Division and supports the above strategic objective by providing end-to-end capabilities to deliver data and information products to users. The data resulting from NASA's missions are a valuable resource that needs to be preserved for the benefit of future generations. These observations are the primary record of the Earth's environment and therefore are the key to understanding how conditions in the future will compare to conditions today. In the near-term, as long as the missions' data are being used actively for scientific research, it continues to be important to provide easy access to data and services commensurate with current information technology. For the longer term, when the focus of the research community shifts toward new missions and observations, it is essential to preserve the previous mission data and the information. This will enable a new user in the future to understand how the data were used for deriving information, knowledge and policy recommendations and to "repeat the experiment" to ascertain the validity and possible limitations of conclusions reached in the past and to provide confidence in long term trends that depended on data from multiple missions. While NASA is not legislatively mandated to preserve data permanently as other agencies such as are the USGS, NOAA and NARA, it is essential for NASA to preserve all the data and associated content beyond the lives of NASA's missions to meet NASA's near-term objective of providing access to data and services for active scientific research. Also NASA has to ensure that the data and associated content are preserved for transition to permanent archival agencies. To fulfill this responsibility, identification of the specific content items that need to be preserved from each of NASA's missions is essential. The purpose of this document is to specify the content items. This document focuses on the "what" (i.e., the content) and not the "how" (i.e. representation of content).

Specifications for preservation information content complement existing archive standards. NASA and the international Consultative Committee for Space Data Systems (CCSDS) member space agencies have long recognized the importance of developing information standards for use in long-term preservation of space-related data collections. Volunteers have developed recommendations titled the Reference Model for Open Archival Information System (RM-OAIS). Subsequent activities continue to expand through a range of related interests that reach toward more practical guidance for developing agency standards. They include provider-archive interchange recommendations (2004) and packaging of data and metadata (XFDU), to facilitate information transfer and archiving (2008). An updated version of the Reference Model (RM) is under review (2009). The CCSDS also is developing ISO 16363 that specifies requirements for certification of trustworthy digital repositories, based on the OAIS Reference Model, and ISO 16919 that describes how to audit archives for compliance with the requirements. The CCSDS/ISO recommendations that most closely relate to a specification for preservation information content are termed Representative Information needed for a full understanding of the digital data objects.

The Reference Model for Open Archival Information System and related work by CCSDS does not provide guidance on the specific types of information that should be preserved with Earth science observational data. However, the Reference Model does give OAIS-compliant archives ground rules and guidance in several important areas. An OAIS-compliant archive should:

- Negotiate for and accept appropriate information from information producers.
- Obtain sufficient control of the information provided to the level needed to ensure Long-Term Preservation.
- Determine, either by itself or in conjunction with other parties, which communities should become the Designated Community and, therefore, should be able to understand the information provided.
- Ensure that the information to be preserved is independently understandable to the Designated Community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information.
- Make the preserved information available to the Designated Community.

These guiding principles will help in developing standard representative information requirements for Earth Science data.

At this time, there are no standards that address content to be preserved for the benefit of future Earth science investigations, especially for future long time series climatological studies. A common approach and consistency across organizations (national and international) and scientific disciplines would ensure that future long-term archives preserve necessary content, particularly since data needed for long-term science studies come from multiple organizations and disciplinary areas. Coordination among organizations will be needed to arrive at common standards. Adoption of this document as NASA's specification of content for preserving Earth science data will be a first step in such coordination. It is possible that this specification will need to be updated as a broader standard is developed, but this specification should serve the interim purpose of gathering preservation content from past and current missions and of ensuring that future mission planning includes preservation of the content items needed in the long-term.

The content items identified in this document are based on:

- The US Global Change Research Program (USGCRP, 1998) Workshop on Global Change Science Requirements for Long-Term Archiving – October 28-30, 1998, Boulder, sponsored jointly by NASA and NOAA;
- Recent work by the ESDIS Project with several science teams whose instruments approaching the end of their lives; and
- Recent work by ESDIS Project staff as participants in the Earth Science Information Partner (ESIP) Federation's Data Preservation and Stewardship Cluster (ESIP, 2011).

The USGCRP workshop had participants representing a wide range of scientific disciplines. The participants developed a number of use cases, considering cases where:

- Existence of a data archive allowed reprocessing to produce new products for global change research;
- Existence of a data archive allowed pursuit of previously unanticipated applications;
- Lack of fully comprehensive data archives severely limited the use of data for scientific research; and
- Scientific questions and hypotheses required long-term archive services.

Based on these use cases and discussions at the workshop, a number of content items were identified as important for preservation.

The ESDIS Project staff has worked with the EP-TOMS, ICESat GLAS and Aura HIRDLS instrument teams to identify the types of information that these teams consider important for preservation in addition to their raw data and derived products that are already in one of the EOSDIS DAACs. The discussions with these instrument teams included the items identified in the USGCRP report as a starting point.

The ESDIS Project staff has been active in the ESIP Federation's Data Preservation and Stewardship Cluster in proposing and developing an emerging standard for Provenance and Context Content. Representatives from several U.S. agencies are involved in this cluster, including NOAA and USGS. While developing such a standard and having it approved through an international standards body is a prolonged process (generally a few years), the content items identified so far are sufficient to provide a preliminary version of a specification for use within NASA.

The content specification is organized into the following eight requirements: Preflight/Pre-Operations Calibration, Products (Data), Product Documentation, Mission Calibration, Product Software, Algorithm Input, Validation and Software Tools. Each of these is described in turn in the following eight sections along with the rationale for why each of the identified items is needed. The description of each specification requirement is expected to evolve over time. The ESDIS Project configuration management process will be used to manage changes.

The focus of this document is intentionally on the content (i.e., the "what") of items that must be preserved rather than on the method or representation (i.e., the "how"). Some variation is expected in the degree with which these requirements are met depending on the phase of the mission at which the requirements are addressed. When applied to older missions where the

projects have ended and the principals are not accessible, these requirements may not be satisfied fully while missions that are still in planning should be able to meet the requirements fully. The details of implementation shall be worked out while developing Interface Control Documents between the mission Project and the NASA ESD-assigned Data Center, which shall be held under configuration control by the ESDIS Project.

CONTENT SPECIFICATIONS

1 Preflight/Pre-Operations Calibration

1.1 Instrument Description

- Item Description: Documentation of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.). Components of documentation include: instrument specifications, vendor calibration reports, user guides, operations concepts and data acquisition timeline, spectral and radiometric calibration reports. (It is recognized that some of the items in this category may be proprietary or International Traffic in Arms Regulations (ITAR) sensitive. When providing such items to the Data Centers, the mission Projects should include any distribution restrictions and expiration dates for those restrictions.)
- Rationale: Needed for users to understand how the instrument operates. Documentation of measurements made before deploying instruments in space (or *in situ*) will help establish a baseline and help users understand changes that may have occurred over time while in operation.

1.2 Preflight/Pre-operational Calibration Data

- Item Description: Numeric (digital data) files of Instrument/sensor characteristics including pre-flight or pre-operational performance measurements (e.g., spectral response, instrument geometric calibration (geo-location offsets), noise characteristics, etc.).
- Rationale: Measurements made before deploying instruments in space (or *in situ*) will help establish a baseline and help users understand changes that may have occurred over time while in operation.

2 Products (Data)

2.1 Raw Data and Level 0 through 4 Products

- Item Description: Raw data are data as measured by a spaceborne, airborne or *in situ* instrument. Product levels 0 through 4 are defined at http://science.nasa.gov/earth-science/earth-science-data/data-processing-levels-for-eosdis-data-products/
- Rationale: Preservation of raw data, Level 0 data or Level 1a products is required for regeneration of any higher-level products in case errors are discovered or better atmospheric transmission/absorption/reflectance or scattering models become available in the future. It is important to preserve either the means of regenerating the higher-level

products or the products themselves to ensure reproducibility and verifiability of scientific results.

2.2 Metadata

- Item Description: Information about data to facilitate discovery, search, access, understanding and usage associated with each of the data products. Metadata shall follow standards described in "Metadata Requirements – Base Reference for NASA Earth Science Data Products" (NASA, 2011).
- Rationale: Data cannot be located or obtained without discovery, search and access metadata. Data cannot be used without metadata needed for understanding and usage. For example, in some production environments it becomes critical that the granule-level metadata indicates which version of software was used to produce a particular granule. It is also important to include provenance information in granule level metadata to enable users to determine all inputs and parameters used in generating the granule.

3 Product Documentation

3.1 Product Team

- Item Description: Names of science team leads and product team members (development, help desk and operations), roles, performing organization, contact information, sponsoring agencies or organizations and comments about the products. As responsibility changes hands over time, the names of individuals and periods during which they were responsible for various aspects of the product should be documented.
- Rationale: It is important to know who were responsible for the products so that they are appropriately credited. Even if specific individuals are not available in the future to provide personal knowledge, their roles and responsibilities may be informative about product quality/validity and consistency, and their publications may provide relevant insights. It is important to capture from the product team any comments about the products before the individuals move on to other activities and become unavailable to provide help to users.

3.2 Product Requirements and Designs

- Item Description: Requirements and designs for each product, either explicitly or by reference to the requirements/design documents. Product requirements and designs should include content, format, latency, accuracy and quality.
- Rationale: Explains expectation about products when the project was initiated. Useful to compare with what was actually accomplished (as recorded in validation documents)

3.3 Processing and Algorithm Version History

- Item Description: For all products held in the archive, documentation of processing history and production version history, indicating which versions were used when, why different versions came about, and what the improvements were from version to version. For all products held in the archive, the versions of source code used to produce the products should be available at the archive. Granule level metadata should indicate which version of software was used for producing a given granule. In the case of some datasets

all versions of products may be maintained. In other cases, only the latest and penultimate versions may be maintained, with some samples of product granules of each of the historical versions. In the case where different versions of ancillary, input data, or calibration were used, the history of those changes should be available as part of the processing history.

- Rationale: It is important to maintain at least the history of all the versions and the rationale for changes in order to preserve the scientific record. Traceability of inputs as well as methods that were used in generating product granules that were used in scientific publications is essential to the scientific method.

3.4 Product Generation Algorithms

- Item Description: Detailed discussion of processing algorithms, outputs, error budgets and limitations with suggested level of detail given below:
    - Processing algorithms and their theoretical (scientific and mathematical) basis, including complete description of any sampling or mapping algorithm used in creation of the product (e.g., contained in peer-reviewed papers, in some cases supplemented by thematic information introducing the data set or derived product) - geo-location, radiometric calibration, geophysical parameters, sampling or mapping algorithms used in creation of the product, algorithm software documentation, & high-level data flow diagrams.
    - Description of how the algorithm is numerically implemented, including possible issues with computationally intensive operations (e.g., large matrix inversions, truncation and rounding).
    - Description of the output data products at a level of detail to determine if the product met specified product requirements.
    - Description of all assumptions that have been made concerning the algorithm performance estimates and any limitations that apply to the algorithms (e.g., conditions where retrievals cannot be made or where performance may be significantly degraded.)
    - Discussion of various error estimates and the error budget.
- Rationale: In order for any product to be used in a scientifically valid manner, it is important to document the theoretical basis for the algorithms used to generate it and the limitations if any. The above documentation should be available and updated for each version of product delivered to the Data Centers so that users of a particular version know exactly how the version was generated.

3.5 Product Quality

- Item Description: Description of the impact to product quality due to issues with computationally intensive operations (e.g., large matrix inversions, truncation and rounding). Documentation of product quality assessment (methods used, assessment summaries for each version of the datasets). Description of embedded data at the granule level including quality flags, product data uncertainty fields, data issues logs, etc. Relevant test reports, reviews, and appraisals. Flowed-through effects of sensor noise,

calibration errors, spatial and spectral errors, and/or un-modeled or neglected geophysical phenomena on the quality of products. Description of potential future enhancements to the algorithm, the limitations they will mitigate, and other useful related information and links.

- Rationale: Users need to understand the known caveats associated with products to ensure their proper usage. It will be helpful to document potential improvements to algorithms that (for whatever reason) were not possible to implement for the archived products.

3.6 Product Application

- Item Description: Useful references to published articles about the use of the data and user feedback received by the science and instrument teams about the products. Includes reports of any peculiarities or notable features observed in the products.
- Rationale: Provides additional help in understanding usage of data products besides the algorithm description and source code. History of users' assessments would be useful for understanding any issues with the products.

4 Mission Calibration

- Item Description: Depending on the instrument and type of platform, the appropriate descriptions of instrument/sensor radiometric and geometric calibration methods, and noise characteristics as well as the data files associated with calibration and the source code used in applying the calibration algorithms. Documentation of in-line changes to calibration or to instrument or platform operations or conditions that occur throughout the mission e.g., instrument events and maneuvers; attitude and ephemeris; aircraft position; event logs (including data gaps); calibration look-up tables; and any significant external events that may have impact on the observations. In the case of in situ data, station location and any changes in location, instrumentation, controlling agency, surrounding land use and other factors which could influence the long-term record.
- Rationale: Documentation of calibration as the instrument characteristics change over time are important to be able to use data in a meaningful way, and to be able to generate long-term time series ensuring consistency over time. Records of instrument operations history, environment and any mission anomalies are important to understand any quirks in data products. Source code will help users in understanding how the calibrated products are generated. It may be possible to use source code (under the right circumstances) to re-implement calibration software. All versions of software should be preserved and relation between product versions and software versions should be documented.

5 Product Software

- Item Description: Product generation software and software documentation. Source code used to generate products at all levels. Software release notes, including references to commercial software libraries and operating system routines used in the code. Descriptions of data products' structure, format, range of values and special fill or error

values. All information needed to verify what output data was created by a run, including data volume and file sizes; i.e. to verify that all expected datasets were produced in the expected format. Documentation that lists the complete set of expected exceptions, and describes how they are identified, trapped, and handled. Documentation needs to identify the source for values of constants and look-up tables used in the algorithm, or explain how they were calculated. The following should be included if a case can be made for future use: Description of all test plans that were produced during development, including references to the artifacts. Descriptions of data sets used for software verification and validation, including unit tests and system test, either explicitly or by reference to the developer's test plans. Test reports or summary of the test results in sufficient detail to indicate that the products met requirements.

- Rationale: Product software source code and production rules provide the definitive procedural steps that document the exact implementation of the algorithm as described in algorithm theoretical basis documents. Product software information documents the relation between product versions and software versions. Product software is needed when considering use of the mission collection in long multi-mission time series to understand procedural impacts relative to other instrument algorithm implementations. When examining local physical artifacts in a mission collection (spatial or temporal), product software provides a way for users to know how a particular geophysical value in the product or product metadata was derived from the combination of inputs. The product software will enable users to know when and how extreme values or unacceptable observations were flagged and treated (e.g. not included) in a particular derived geophysical or metadata value. The product software will help users identify the source contributions to errors and uncertainties of a particular observation. Earlier versions of software should be preserved when used to generate a version of the product that was available to the community and resulted in cornerstone findings (as advised by science community representatives).

6 Algorithm Inputs

- Item Description: Identify all ancillary data or other data sets used in generation or calibration of the data or derived product at all levels. Ancillary data should be stored with the products unless it is available from another permanent archive facility. Include the name and location of the ancillary data archive facility if ancillary data will not be stored with the products. Complete information on any ancillary data or other data sets used in generation or calibration of the data set or derived product, either explicitly in data descriptions or by reference to appropriate publications. Information should include full description of the input data and attributes covering all input data used by the algorithm, including primary sensor data, ancillary data, forward models (e.g. radiative transfer models, spectral line-lists, optical models, or other model that relates sensor observables to geophysical phenomena) and look-up tables. At granule level, include information on all inputs (including ancillary or other data granules, calibration files, look-up tables, ground control, climatology etc.) that were used to generate the product. At the appropriate level (granule or dataset) include calibration parameters, precision orbit & attitude data; climatological norms, geophysical masks; First-guess fields from

numerical weather or climate models; spectrum and transmittance information. Describe any important programming and procedural aspects related to implementing the algorithm into operating code.

- Rationale: The algorithm input information is needed by users investigating the products for long multi-mission time series. Investigators need this information to understand the relative contributions of each input to an output geophysical value in the product, both at a global scale and across the life of the mission, and at local spatial (e.g., regional focus) and temporal (e.g., extreme event focus) scales. Similarly, when investigating a local physical artifact in the mission collection (e.g., regional or extreme event), the algorithm input provides a way for users to see whether the artifact is present in ancillary data such as the first guess field, or in climate fields versus from the instrument observations. This is especially important when investigators want to consider the impact of new improved ancillary values or ancillary geophysical relationships such as land-ocean masks or standard atmosphere profiles could impact derived climate trends, significantly reduce error or bias in a derived product. Knowledge of all algorithm inputs is critical for assessing repeatability and usability of the experiment's results.

7 Validation

- Item Description: Datasets and documentation. Accuracy of products, as measured by validation testing, and compared to accuracy requirements. Description of validation process, including identification of validation data sets, measurement protocols, data collection, analysis and accuracy reporting. This should include a description of Cal/Val plans & status, as well as a detailed history of validation activities and validation data sets along with metadata from previous validation exercises.
- Rationale: Users will need to understand the procedures used for validation during the mission lifetime, as well as caveats associated with products to ensure their proper usage. Investigators need evidence of the observed geophysical references for comparing calibrated and derived geophysical values to other long-term observational data sets. This evidence is especially important for satellite based observations because the validation studies are often limited to comparisons with in-situ or aircraft observations from regional campaigns for finite time periods.

8 Software Tools

- Item Description: Product access (reader) tools. Software source code that would facilitate use of the calibration data, ancillary data and the data products at all levels. Includes software source code useful for creating programs that will read and display the calibration data, ancillary data and product data and metadata values. Commercial tools should be identified with appropriate references. Include release notes, identify sample input and show the corresponding output results.
- Rationale: Software tools help facilitate use of data and metadata as well as confirm documentation of the data and metadata structure. Provides an example of the data and metadata values users should expect to see from the products.

REFERENCES

CCSDS, 2002: *Reference Model for an Open Archival Information System (OAIS)*. Recommendation for Space Data System Standards, CCSDS 650.0-B-1. Blue Book. Issue 1. Washington, D.C.: CCSDS, January 2002. [Equivalent to ISO 14721:2003.]

CCSDS, 2009: Audit and Certification of Trustworthy Digital Repositories, Draft Recommended Practice. CCSDS 652.0-R-1. Issue 1, Washington, DC: CCSDS, October 2009. http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206520R1/Attachments/652x0r1.pdf

CCSDS, 2010: Requirements for Bodies Providing Audit and Certification of Trustworthy Digital Repositories, Draft Recommended Practice. CCSDS 652.1-R-1. Issue 1, Washington, DC: CCSDS, October 2010 http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206521R1/Attachments/652x1r1.pdf

ESIP, 2011: Preservation and Stewardship Cluster of the Earth Science Information Partners Federation, http://wiki.esipfed.org/index.php/Preservation_and_Stewardship.

NASA, 2011: Metadata Requirements – Base Reference for NASA Earth Science Data Products, September 2011.

USGCRP, 1998: *Global Change Science Requirements for Long-Term Archiving*, Report of the Workshop, October 28-30, 1998, National Center for Atmospheric Research, Boulder, CO, Sponsored by NASA and NOAA, through the USGCRP Program Office.